Published online July 7, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 12 ***3581–3589***
doi:10.1093/nar/gkh681

New strategy for the representation and the integration of biomolecular knowledge at a cellular scale

Roland Barriot[1,2], Jérôme Poix[3], Alexis Groppi[1], Aurélien Barré[1], Nicolas Goffard[1], David Sherman[1,2], Isabelle Dutour[1,2] and Antoine de Daruvar[1,*]

[1]Centre de Bioinformatique de Bordeaux, Université V. Segalen Bordeaux 2, 146 rue Léo Saignat, 33076 Bordeaux, France, [2]LaBRI, Laboratoire Bordelais de Recherche en Informatique, UMR CNRS 5800, 351 cours de la Libération, 33405 Talence Cedex, France and [3]Laboratoire Statistique Mathématique et ses Applications, EA 2961, Université V. Segalen Bordeaux 2, 146 rue Léo Saignat, 33076 Bordeaux, France

Received March 8, 2004; Revised and Accepted June 4, 2004

ABSTRACT

The combination of sequencing and post-sequencing experimental approaches produces huge collections of data that are highly heterogeneous both in structure and in semantics. We propose a new strategy for the integration of such data. This strategy uses structured sets of sequences as a unified representation of biological information and defines a probabilistic measure of similarity between the sets. Sets can be composed of sequences that are known to have a biological relationship (e.g. proteins involved in a complex or a pathway) or that share similar values for a particular attribute (e.g. expression profile). We have developed a software, BlastSets, which implements this strategy. It exploits a database where the sets derived from diverse biological information can be deposited using a standard XML format. For a given query set, BlastSets returns target sets found in the database whose similarity to the query is statistically significant. The tool allowed us to automatically identify verified relationships between correlated expression profiles and biological pathways using publicly available data for *Saccharomyces cerevisiae*. It was also used to retrieve the members of a complex (ribosome) based on the mining of expression profiles. These first results validate the relevance of the strategy and demonstrate the promising potential of BlastSets.

INTRODUCTION

Cellular functions result from molecular mechanisms that are individually studied using a combination of sequencing and post-sequencing experimental approaches. Understanding the tight coupling between these mechanisms at a cellular scale requires efficient methods and tools for integrating a huge collection of highly heterogeneous data. These data, diverse both in structure and in semantics, include functional and structural sequence annotations, expression profiles of genes and proteins, molecular interactions between biomolecules, etc. Integrating these data leads to a better understanding of cell-wide processes and ultimately contributes to greater knowledge of the organization and functioning of the cell.

Frequently used in bioinformatics, the concept of 'data integration' is imprecise and refers to several concepts.

- Integration through linking. In the context of systems such as SRS (1) or ENTREZ (2), which provide means for querying and navigating in multiple databanks, integration means the exploitation of links that are, in most cases, cross references between entries from different databases. This integration is very general in scope, as it allows navigation through a wide and complex network of links across heterogeneous data, but it is also limited in functionality because it only uses static relationships between individual entries.
- Integration through modeling. In 'Systems Biology' (3,4), integration means using a formal language to specify dynamic models of biological processes. In this context, the integration is usually restricted to dynamic relations between certain types of data. However, such an integration is functionally ambitious as it aims at revealing new knowledge through the emergent properties of the model.

Between 'linking' and 'modeling', the concept of neighborhood was proposed in 1998 by Danchin (5). Danchin stated that the availability of complete genomes and proteomes offers the opportunity to move the focus from individual biological objects such as genes or proteins, to the relationships (neighborhoods) between these biological objects. As an illustration of the concept, the approach was used to design a genome viewer named Indigo (6). Indigo provided a set of graphical views, which offered a visualization of different types of putative relationships among genes: physical proximity on the chromosome, co-citation in the literature, similar usage of the genetic code, etc. The user of Indigo could visually identify similarities and thus correlations between different neighborhoods. The integration capabilities of Indigo, while innovative, suffered from severe limitations: the application did not rely on a standard data structure, each relationship was 'hard

[*]To whom correspondence should be addressed. Tel: +33 5 57 57 12 47; Fax: +33 5 57 57 12 47; Email: antoine.daruvar@pmtg.u-bordeaux2.fr

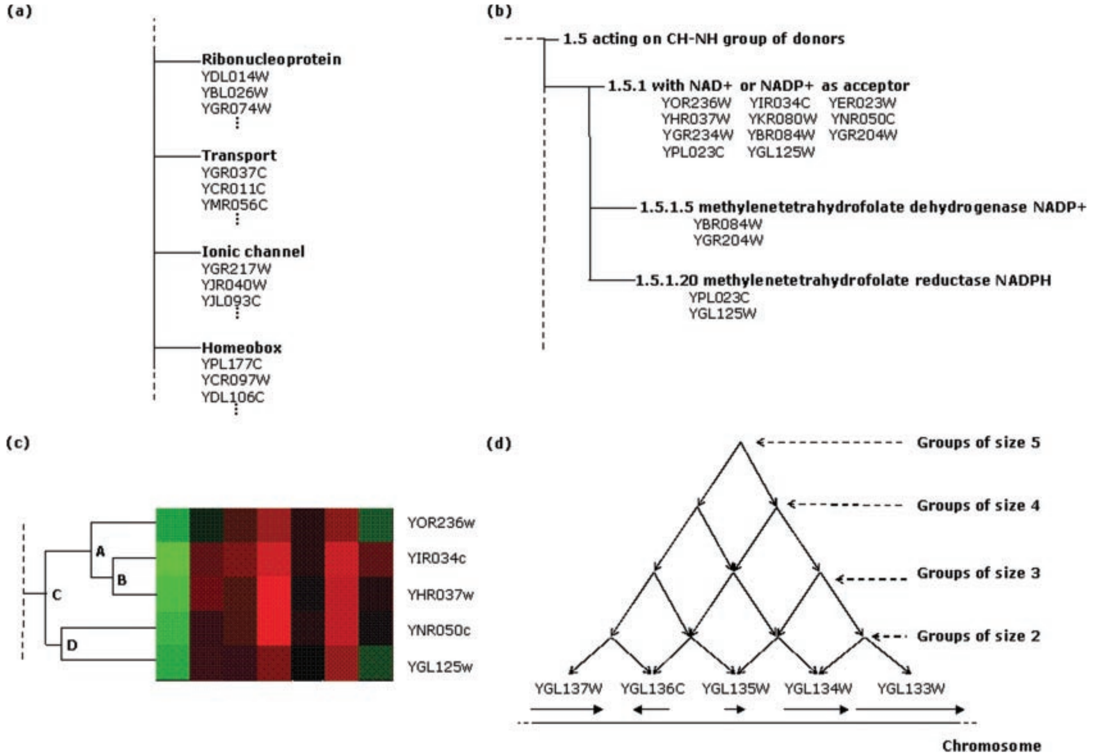


Figure 1. Examples of set definitions for biological information in yeast (genes are identified using the systematic nomenclature). (**a**) SWISSPROT keywords: a set contains all the sequences that are annotated with a given keyword. The sets are independent of each other and form a star graph. (**b**) Enzyme EC Numbers: each class in the hierarchical classification of enzymes is used to define a set. This set contains all sequences that are annotated as being part of the class plus all the sequences attached to the corresponding sub-classes. i.e.: the set of class 1.5.1 contains all the proteins in its sub-classes (sub-classes containing only one protein are not represented here). The sets are hierarchically organized and form a tree. (**c**) Expression profiles: each node of the binary tree resulting from a hierarchical clustering of expression profiles is used to define a set. This set contains all sequences from the corresponding branch. The sets are hierarchically organized and form a binary tree. (**d**) Chromosomal localization: a set is defined for each node of an implicit lattice structure built on top of the chromosomal localization of genes. All possible sets of adjacent genes are thus defined: from pairs to the complete chromosome. The sets are hierarchically organized and form a Directed Acyclic Graph (DAG).

coded' in flat files in a given format, and no measure of similarity between relationships was offered.

We present here a new strategy for data integration, inspired by Danchin's neighborhood concept. Our approach makes it possible to dynamically bring together heterogeneous information available at the scale of complete genomes or proteomes. It allows integration of broad datasets and aims at revealing new correspondences between them.

The basic principles of this strategy are:

- use of sets of biological sequences (genes, gene products, etc.) as a unified data structure;
- systematical conversion of available biological knowledge into sets of sequences;
- storage of these sets in a database which supports a standard import format;
- use of a probabilistic model to measure the similarity between sets.

This strategy was implemented in BlastSets, a software which allows the user to submit one or several sets in order to retrieve from the database all the sets that are found to be significantly similar. A Web interface has been developed and is publicly accessible (http://cbi.labri.fr/outils/BlastSets/).

Although, several systems (7,8) use a measure of similarity between sets of sequences defined on various criteria, we are the first to propose to use this approach as a general solution for the integration of bio-molecular data at the scale of the cell.

In the following, we explain the method in detail and we present different uses of BlastSets which illustrate the relevance and the power of the approach.

METHODS

In this section, we explain how sets of sequences can be defined to capture different types of biological information and how such sets are organized to form 'BlastSets Classifications'. We then explain the mathematical methods used to perform set comparisons in order to identify significant similarities.

Defining sets of sequences

Schematically, sequences from an organism are grouped to form a set when they share identical or similar values for a particular attribute. Those attributes are defined using diverse sources of biological information (see Figure 1 for detailed examples):

- cellular process: set of genes which contribute to the same biological pathway

- genome features: set of genes' neighbors on the chromosome(s)
- experimental results: set of genes with similar expression profiles
- physico-chemical property: set of proteins having similar isoelectric points
- sequence annotation: set of proteins that share a common keyword

We only consider genes coding for proteins and we do not distinguish between genes and gene products in the set definition. This allows us to bring together information relative to genes or to proteins.

Depending on the biological criteria, the definition of the sets requires more or less processing. In a number of cases, the set definition is rather straightforward:

- When sequences are found to form together a biological object, the grouping of sequences is obvious: i.e. each protein complex purified and identified by Cellzome (9) defines a set.
- Sequence attributes that correspond to discrete values, such as a keyword or a structural domain, can be easily used to define sets (Figure 1a): to each keyword or each domain corresponds the set of all the sequences that share this attribute.
- Systematic classifications provide collections of sets: e.g. EC Numbers (10) classify enzymes in different classes and subclasses. We can see each class as a set of sequences having the same type of enzymatic activity as shown in Figure 1b.

In the above examples, sets can be derived directly from the original data. However, for other criteria, rules must be defined to create the sets.

- One challenging case is when the criterion is a measure of continuous values. If we look at the position on the chromosome of a coding sequence, we have a range of possible values. In order to build sets, one can define a window of a certain size, then slide this window along the chromosome and build the resulting sets of adjacent genes. One problem is that we do not know a priori the relevant window size to use. The choice of the window size implies a choice in the level of granularity used to aggregate adjacent genes. We can choose to look at very small sets of contiguous genes like pairs or we can consider large segments of chromosomes. The diversity of the size of known operons shows that a fixed size is not appropriate. In order to keep sets that cover all possible levels of granularity from pairs to complete chromosome, we propose a hierarchical aggregation of neighboring genes as shown in Figure 1d.
- In order to derive sets from expression profile data, we choose to rely on the hierarchical clustering (11) of the profiles. This is one of the well-established methods for analyzing these profiles. It results in a binary tree. Here again, there is no clear rule that can be applied to retrieve from the binary tree the sets of sequences that are significantly co-regulated, and thus correspond to real biological signals. In order to capture as much information as possible from this tree, we retrieve and store the sets attached to all nodes corresponding to all granularities in gene aggregations (Figure 1c).

The first step in our data integration strategy is to systematically build sets as described above. The goal of this process is to project into a unified data structure the largest possible fraction of biological knowledge at the molecular and cellular level.

BlastSets Classifications

Sets, as defined above, are not independent entities: a set belongs to a collection of sets that is derived from a particular biological criterion. Furthermore, there exist relationships, typically inclusion, between sets of a given collection. Those relationships can be described in various types of graphs: i.e. simple star graph (Figure 1a), trees (Figure 1b and c) or lattice (Figure 1d). Actually, all the graphs that can be required to describe the relationships among sets are Directed Acyclic Graphs (DAG).

In order to keep track of these relationships, which cannot be represented at the level of individual sets of sequences, we introduced a new structure: 'BlastSets Classification'.

Definition: A BlastSets Classification is made of three elements:

- a biological criterion;
- a collection of sets of sequences;
- a DAG that describes relationships among the sets: in this DAG, the nodes correspond to sets and the edges represent the inclusion of the sequences of a node (target) into another (source).

Set comparisons

In this section, we discuss only the principles of set comparisons and score significance. Full details on the mathematics involved are provided as supplementary material online (http://cbi.labri.fr/outils/data/blastsets/).

We use the hypergeometric distribution to compare two sets of sequences and measure a 'distance' between them. This method was used by (7,8,12) and proved to be simple and efficient. We formulate this measure as the probability of having at least the observed number of sequences in common between two sets, that may differ in size, built over all the possible sequences. We denote this probability as the P-value.

A P-value is considered significant (sets are similar) if it is less than or equal to a certain threshold. Multiple comparisons are performed as a set is compared to all the sets of a BlastSets Classification. Moreover, the sets' compositions are not independent within a BlastSets Classification. Thus, we want to adjust the P-value significance threshold to the considered target sets. In (7,12), a Bonferroni correction is used. This correction considers only the number of tests conducted and not the actual target sets composition. In order to adapt the cut-off to the considered target sets, we use the probability distribution of the minimum P-values.

The minimum P-values probability distribution function is defined for a BlastSets Classification and a query set size. For a given P-value, it gives us the probability of obtaining a P-value at least as good by submitting a random set of the same size. Unfortunately, it is practically impossible to compute this function, so we approximate it by sampling to obtain an empirical function. Thousand Monte Carlo simulations were used by (8) to build an empirical function. Further investigations and a proposition by Dufour (13) allow us to perform only 500 simulations without loosing much precision (~1%). The obtained empirical function provides an *estimation* of the

Table 1. BlastSets Classifications available in the database

Species	Classification type	Name	No. of sets
S.cerevisiae	Systematic classification	KEGG metabolic pathways	131
		Enzyme EC numbers	280
		Funcat from the MIPS	175
		Subcell from the MIPS	44
		GeneOntology—molecular function	791
		GeneOntology—biological process	1053
		GeneOntology—cellular component	303
	Expression data	27 microarray datasets from the Stanford Microarray Database (SMD)	150760
	Physical interaction	Cellzome proteic complexes	226
	Keywords	SWISSPROT Keywords	277
	Total		154040
E.coli	Systematic classification	KEGG metabolic pathways	148
		Enzyme EC numbers	274
	Expression data	9 microarray datasets from the Stanford Microarray Database	36511
	Keywords	SWISSPROT Keywords	277
	Total		37210
B.subtilis	Systematic classification	Enzyme EC numbers	248
		KEGG metabolic pathways	146
		GeneOntology—molecular function	263
		GeneOntology—biological process	298
		GeneOntology—cellular component	31
	Expression data	2 microarray datasets from the Stanford Microarray Database	11168
	Total		12154

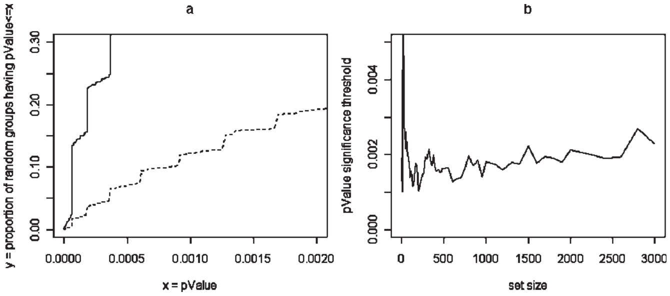


Figure 2. *P*-value significance determination by the mean of empirical probability distribution function of the minimum *P*-values (**a**) The empirical distribution functions of the minimum *P*-values for two BlastSets Classifications for a query of size 50. The solid line corresponds to a hierarchical clustering of expression profiles, and the dashed line corresponds to the GeneOntology molecular function branch. For a cut-off of 0.1 and a query of size 50, the *P*-value threshold for significance is $8.1E^{-4}$ for the Gene Ontology whereas it is $6.2E^{-5}$ for the transcriptome experiment. (**b**) Estimated significant *P*-value threshold depending on the query set size for a cut-off of 0.1 for the Cellzome BlastSets Classification.

probability of obtaining a *P*-value at least as good, so we denote it the *E*-value. The *E*-value provides a measure of significance of a *P*-value (the lower, the better).

From one BlastSets Classification to another, the number of sets varies (see Table 1) as well as the distribution of the sizes of the sets. The empirical function is thus computed for each BlastSets Classification separately. Figure 2a illustrates that for a given cut-off (*E*-value upper bound), the *P*-value significance threshold depends on the BlastSets Classification.

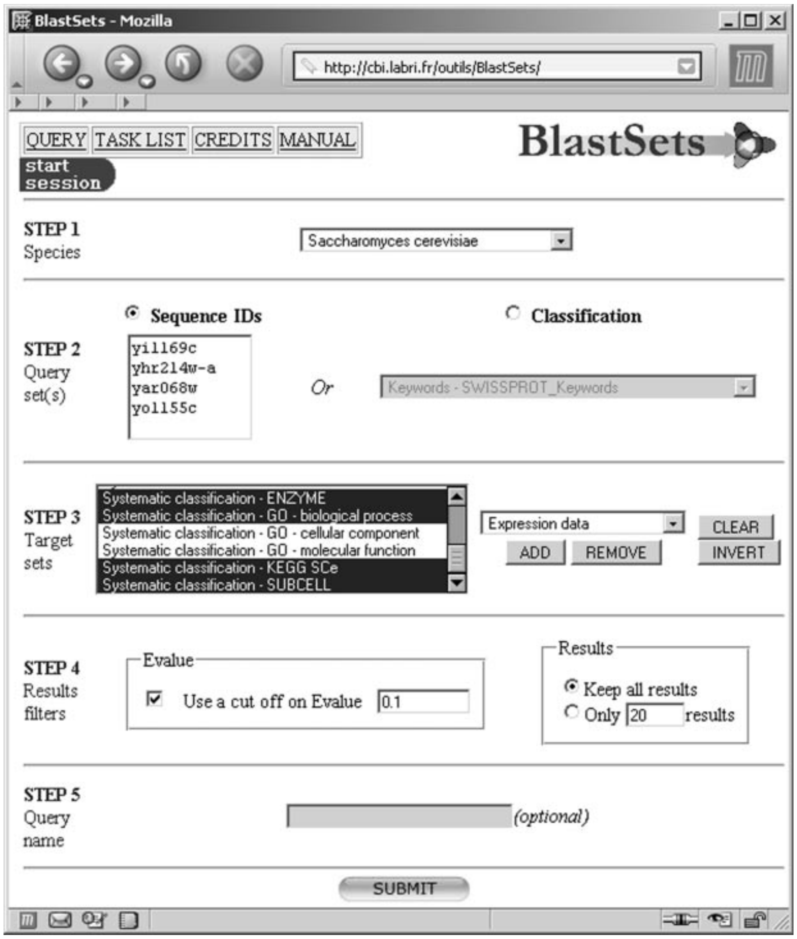
The empirical function must be computed for all possible sizes of query sets. Figure 2b shows that the *P*-value threshold for a given cut-off (here the *E*-value upper bound is set to 0.1) depends on the query set size. In this context, a Bonferroni correction, which gives a constant *P*-value threshold depending only on the number of target sets, does not appear to be appropriate.

SYSTEM

We have designed and developed a system that accepts queries and returns a list of significantly similar sets. A query is made of one or more query sets to be compared to a selection of one or more target BlastSets Classifications. A query set can be specified by submitting a list of sequence identifiers, or by choosing a BlastSets Classification. In the latter case, each set of the chosen BlastSets Classification serves as a query set. To handle the multiplicity of aliases that are used to refer to each gene or protein, we use AliasServer (14). This application provides services for equivalent identifier conversions. The sequence identifiers submitted by the user are automatically converted to the identifier used internally in the BlastSets database: a checksum key computed from the sequence using the CRC64 algorithm.

The result of a query is a list of hits. A *hit* corresponds to a significant similarity between a query set and a target set and contains the following information:

- If the query sets comes from a BlastSets Classification, then details on the node corresponding to the query set are given (BlastSets Classification, node name, short description, number of sequences).
- Details on the target set (BlastSets Classification, node name, short description, number of sequences).
- The number of sequences in common between the query and the target sets.
- The *P*-value, which corresponds to the probability of having at least the observed number of sequences in common between the query and the target sets.
- The *E*-value, which corresponds to the expectation level of having at least as good a *P*-value by comparing a random query set of identical size to the target BlastSets Classification.



STEP 1 Species: Saccharomyces cerevisiae

STEP 2 Query set(s): ● Sequence IDs ○ Classification

```
yil169c
yhr214w-a
yar068w
yol155c
```

Or Keywords - SWISSPROT_Keywords

STEP 3 Target sets:

- Systematic classification - ENZYME
- Systematic classification - GO - biological process
- Systematic classification - GO - cellular component
- Systematic classification - GO - molecular function
- Systematic classification - KEGG SCe
- Systematic classification - SUBCELL

Expression data

STEP 4 Results filters: Evalue ☑ Use a cut off on Evalue 0.1 — Results ● Keep all results ○ Only 20 results

STEP 5 Query name: (optional)

Figure 3. Screenshot of the web interface query tab. Step by step: (step 1) the user has selected the species *S.cerevisiae*, (step 2) pasted a list of four sequence identifiers (step 3) to compare to four BlastSets Classifications (step 4) with a cut-off of 0.1. This web page is publicly accessible at http://cbi.labri.fr/outils/BlastSets/.

A screenshot of the Web interface is shown in Figure 3; it is publicly available at http://cbi.labri.fr/outils/BlastSets/.

To facilitate the loading of new datasets in the database, an XML DTD has been defined for BlastSets Classifications (available at http://cbi.labri.fr/outils/data/blastsets/). The use of such a standard format makes BlastSets an open system: new datasets can be easily added.

As in any complex project involving data manipulation, system consistency and stability may be compromised by the introduction of new data or by small modifications made to the system. To ensure the reproducibility and the reliability of the results, a consistency checking procedure has been developed and is detailed in the online supplementary material (http://cbi.labri.fr/outils/data/blastsets/).

At the time of writing, the database contains 55 BlastSets Classifications concerning *Saccharomyces cerevisiae* (36), *Escherichia coli* (12) and *Bacillus subtilis* (7). The database content is summarized in Table 1.

RESULTS

In order to evaluate the relevance of the BlastSets strategy and method, we tested whether it was able to reproduce results that were obtained by expert annotation of an expression profile experiment. We then tested our tool for its capacity to explore the 'expression profile neighborhood' of genes.

Annotation of an expression profile experiment

We chose the microarray transcriptome analysis by Ferea *et al*. (15), which reports significant altered expression levels for several hundred genes from *S.cerevisiae*. The authors identified four main biological categories for the genes with altered expression: glycolysis, tricarboxylic acid cycle, oxidative phosphorylation and metabolite transport. This result was obtained by manual analysis of the microarrays. Such analyses run up against two major difficulties: the complexity of the task (here 4 experimental conditions and about 6000 genes for *S.cerevisiae*) and the lack of objective measures for assessing the significance of the reported observations.

The raw data (Log$_2$ of Red/Green normalized ratio without any filter) were collected from the Stanford Microarray Database (16). The preclustering file thus obtained was filtered against the list of 5786 validated open reading frames (ORFs) from the GDR Genolevures (17). The hierarchical clustering of the data was done using the Cluster software from Eisen lab (11) with default parameters. The sets corresponding to all the nodes of the resulting binary tree were used to build a BlastSets Classification that was then loaded in the database. It was then used as a query against a BlastSets Classification derived from the KEGG pathways database (18) of *S.cerevisiae* (cf. Table 1). The results obtained automatically by BlastSets are presented in Table 2. They are consistent with those reported in the publication: glycolysis, tricarboxylic acid cycle and oxidative phosphorylation are among the pathways that are found to have most significant hits with expression profile sets.

Interestingly, BlastSets reported some significant hits with additional metabolic pathways (amino acid metabolism and lipid metabolism) that were not mentioned in the original publication (15).

The pathway that was found with the most significant hit, corresponds to translation and the ribosome. This result is not surprising: transcription analyses frequently highlight the strong correlation of expression of the ribosomal proteins. In 2002, Jansen *et al.* (19) while analyzing the relationship between whole genome expression data and protein–protein interaction, found that the subunits of permanent complexes (maintained through most cellular conditions), such as the ribosome and the proteasome, show significant coexpression. In our results, the proteasome was also found to have very significant hits (Table 2).

Mining BlastSets database: exploration of an expression profile neighborhood

In a second experiment, BlastSets was used to explore the expression profile neighborhood of gene sets. A bait set was used as query and compared to BlastSets Classifications derived from a collection of publicly available transcriptome data loaded in the database. At the time of this experiment, the BlastSets database contained 27 different yeast transcriptome datasets (Table 1) from which nearly 150 000 expression sets were derived. As bait sets, we used different samples ranging from 20 to 50 genes randomly selected from a list of 128 validated genes which were assigned to the ribosome in the KEGG database (18). As the different random sets gave very similar results, we arbitrarily present and discuss results obtained with a random set composed of 20 genes.

For each bait set, BlastSets identified a list of significantly similar expression hit sets. Each of the 5640 genes found in at least one hit set was assigned a score, which is its total number of occurrences in all hit sets. Higher scores are expected for genes that show a stronger correlation of expression with the bait set. The genes were sorted by decreasing score. The results are presented in Figure 4.

Roughly speaking, 3 sections can be defined based on the shape of the histogram in Figure 4a: from rank 1 up to 100 are

Table 2. Expression profiles against KEGG pathways in yeast

Rank (on a total of 1533 Hits)	Query node name	Query size	Target node name	Target short description	Target size	Number of IDs in common	*P*-value	*E*-value
1	NODE5526X	1982	13	Translation	210	169	1.56E-44	0
2	NODE5526X	1982	13.1	Ribosome	128	112	1.54E-36	0
3	NODE5490X	376	13.1	Ribosome	128	52	1.86E-29	0
				. . . // . . .				
21	NODE5310X	209	2.1	Oxidative phosphorylation	70	24	5.35E-18	0
				. . . // . . .				
28	NODE5525X	1361	5	Amino Acid Metabolism	223	107	3.47E-16	0
				. . . // . . .				
71	NODE5522X	621	14.7	Proteasome	32	16	2.97E-08	0
				. . . // . . .				
88	NODE5351X	119	6	Metabolism of Other Amino Acids	59	10	2.48E-07	0
				. . . // . . .				
96	NODE5474X	900	1	Carbohydrate Metabolism	171	52	5.21E-07	0
				. . . // . . .				
166	NODE1602X	2	5.2	Alanine and aspartate metabolism	26	2	1.94E-05	0
				. . . // . . .				
168	NODE5505X	1164	5.15	Phenylalanine, tyrosine and tryptophan biosynthesis	23	14	2.15E-05	2.60E-02
				. . . // . . .				
176	NODE5511X	194	3	Lipid Metabolism	49	9	2.85E-05	4.00E-03
				. . . // . . .				
181	NODE2298X	4	1.9	Glyoxylate and dicarboxylate metabolism	14	2	3.25E-05	0
				. . . // . . .				
183	NODE856X	2	1.8	Pyruvate metabolism	34	2	3.35E-05	0
				. . . // . . .				

The table shows a summary of the results obtained when comparing all the (query) sets derived from a hierarchical clustering of an expression profile experiment (15) against (target) sets derived from the KEGG pathways database.

Each line corresponds to a significant hit between a query set and a target set. The hits are sorted by increasing *P*-values. The table only shows the best hits for each different pathway. The . . . // . . . correspond to omitted results that were chosen by applying the following rule (except for the first three rows for illustration): we select hits for which neither the query node nor the target node was already part of a hit with a better rank. The full table (all results) is provided as supplementary material (http://cbi.labri.fr/outils/data/blastsets/).

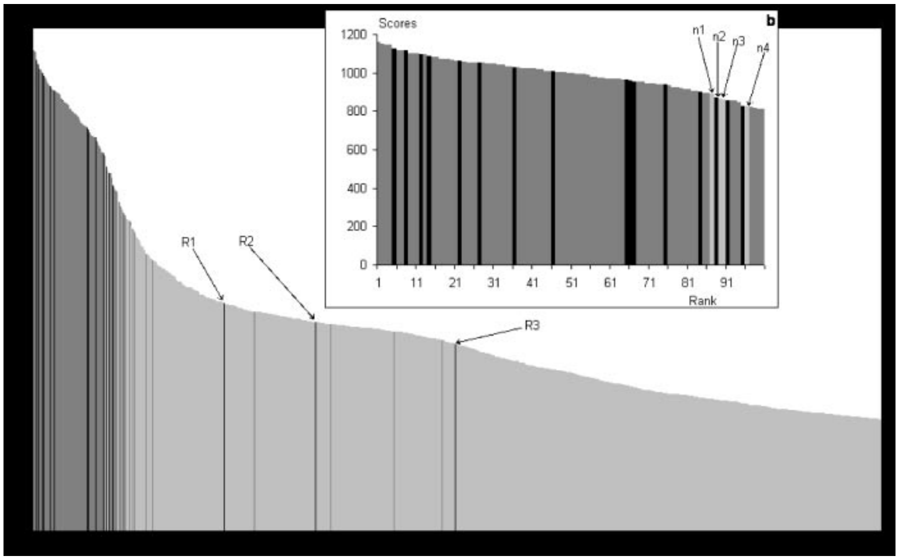


Figure 4. Exploration of expression profile neighborhood using BlastSets A random sample of 20 yeast genes from the KEGG 13.1 pathway (ribosome) was used as the query set to fetch hit (significantly similar) sets among nearly 150 000 sets derived from 27 different yeast transcriptome experiments. Each gene contained in at least one hit set was assigned a score which is its number of occurrences in hit sets. The genes were sorted according to decreasing scores. The genes from the random query set are in black, the other ribosomal genes are in medium gray and genes that are not annotated as members of the ribosome are in light gray. (**a**) Results for gene until rank 1000; (**b**) Results for the genes in the first 100 ranks. n1 (YKL056C), n2 (YMR116C/ACS1), n3 (YNL119W) and n4 (YNL255C/GIS2) represent non-ribosomal genes picked out within the highest scores. R1 (RPS24B/YIL069C), R2 (RLP24/YLR009W) and R3 (RPL22B/YFL034C-A) are ribosomal genes that were present in the random query set and have a rather low score.

found genes with high scores, between rank 100 and 500 the scores are medium and after rank 500 are the low scores. A careful analysis of the results provides a variety of interesting observations.

- Obviously most genes that are annotated as coding for ribosomal proteins (in black and medium gray on Figure 4) are highly concentrated in the high scores section (rank 1 to 100), which confirms both the strong correlation of expression of the ribosomal proteins and the capacity of the BlastSets strategy to efficiently fetch neighbors of a set using a given biological criterion. In addition, the overall order of the genes, especially in the high scores, is very well conserved among the different random samples independent of their size (data not shown).
- The Figure 4b shows a 'zoom' on the first 100 highest scores. 4 non-ribosomal genes are found within these high scores. These genes are YKL056C, YMR116C, YNL119W and YNL255C respectively rank 88, 90, 91 and 97 (n1, n2, n3 and n4 on Figure 4b). The cellular location inferred from direct assay for these four genes is cytoplasmic, thus consistent with participation in the cytoplasmic ribosomal activity. YKL056C, YNL119W are annotated as hypothetical ORFs in *S.cerevisiae*. From our results we predict that the proteins encoded by these genes participate in (or interact with) the translational machinery in *S.cerevisiae*. The third gene (YMR116C—Asc1p) has been identified as a guanine nucleotide binding protein that interacts with the translational machinery in *S.cerevisiae*. It has been described as a ribosome-associated protein with a nearly stoichiometrical association with the ribosome (20). The fourth gene (YNL255C—GIS2) has been described as participating in an intracellular signaling cascade (inferred from genetic interaction). The molecular function inferred from sequence or structural similarity has been putatively assigned to transcription factor activity. Again, our results indicate a possible role closely linked to the ribosome itself or its expression.
- Conversely, there are ribosome-annotated genes that are 'rejected' far from the high scores. Interestingly, this rejection occurs even for genes that were included in the query sample. On Figure 4a, genes R1 (RPS24B/YIL069C), R2 (RLP24/YLR009W) and R3 (RPL22B/YFL034C-A) are ribosome annotated genes that were present in the random query set and have a medium score which indicates a probable weak correlation of expression with the other subunits of the ribosome. In addition, there are 10 ribosomal genes, not visible on Figure 4a, that are rejected in the low or very low scores. These genes are YFR032C-A, YPL249C-A, YMR024W, YMR286W, YBR251W, YLR439W, YDR462W, YNR037C, YJR113C and YGR076C, respectively, ranked 1584, 1653, 1768, 2038, 2376, 2921, 3877, 3885, 4007 and 4303. YFR032C-A (RPL29) was described as a non-essential gene that codes for a 60S ribosomal subunit protein in *S.cerevisiae*. Its deletion leads to a moderate accumulation of half-mer polysomes with little or no change in the amounts of free 60S subunits (21). Its low score indicates that it is probably not co-expressed with other ribosome-annotated genes which corroborates its non-essential role. YPL249C-A (RPL36B) has been described as being part of the 60S large ribosomal subunits (22), and the resulting protein is bound to the 5.8S rRNA (23). Very little data are available on the role of this protein to the ribosome in yeast. Our results demonstrate that the expression of this gene is impaired compared with the other ribosomal genes investigated here. It may be interesting to look very carefully at this gene to clarify its role in the translational machinery of *S.cerevisiae*. All the last eight genes are annotated as being

mitochondrial ribosomal genes. The KEGG 13.1 pathway (ribosome) contains genes both from the cytoplasmic and mitochondrial ribosome complexes. The mitochondrial ribosome clearly has a totally different behavior compared to the expression of the genes composing the cytoplasmic complex. One of these genes (YLR439W) was present in the query sample and despite this, it ends up with a very low score. This demonstrates the robustness of the BlastSets approach.

DISCUSSION

We propose a general strategy for the integration of genomics and functional genomics data. This strategy relies on a unified representation of heterogeneous biological information in the form of sets of sequences and a probabilistic measure of similarity between those sets:

- The sets in the unified representation can correspond to any observed biological relationship among individual sequences: the set of proteins that form a complex, the set of genes that belong to an operon, the set of enzymes that contribute to a given pathway, etc. Sets can also be based on an identical or similar value for a given attribute of the sequences: the set of proteins that share a structural domain, the set of genes physically neighbors on the chromosome, the set of sequences that share a keyword in their annotation, the set of proteins with similar isoelectric points, etc.
- The probabilistic similarity measure is based on the composition of the sets: it uses the hypergeometric law which gives the probability that two sets independently extracted from a population have a certain number of elements in common. This measure allows us to identify correspondences between sets that refer to different biological criteria (i.e. between co-regulated genes and a particular functional class).

We have implemented this strategy in a software system named BlastSets. The kernel of the system is a database where sets are stored together with the corresponding biological information. BlastSets offers the possibility to submit one or several query sets in order to automatically compare them to target sets contained in the database. The best similarities between query and target sets are identified and returned as results. Since good similarity scores (*P*-values) can occur by chance, the *similarity significance* is determined by estimating the probability that a given observed similarity score might be obtained by chance with respect to the database content.

In order to validate our strategy, we used BlastSets to analyze public data on *S.cerevisiae*.

- In a first experiment, the tool was used to annotate results of a transcriptome expression analysis (15). Expression profiles were hierarchically clustered and BlastSets was used to find which pathways, as defined in the KEGG database, corresponded the most to clusters of putatively co-regulated genes. BlastSets automatically identified the pathways that were manually detected by the authors. This first result demonstrates that the approach can be particularly useful and time-saving for the analysis and the annotation of experimental data.
- In the second experiment, we used a random subset of proteins from the ribosome as a query set. BlastSets was used to fetch similar sets from among 150 000 sets corresponding to a hierarchical clustering of 27 expression profile experiments. Most proteins annotated as ribosomal were found to appear with the highest frequencies within the similar sets. Interestingly, the tool was able to retrieve the entire complex starting from an incomplete bait set. In addition, BlastSets identified proteins that were not annotated as ribosomal and that were often found clustered with ribosomal proteins. For those which had no functional annotation, this result strongly suggests their involvement in the translation machinery.

These first results demonstrate that sets of sequences can be used efficiently to represent and integrate heterogeneous biological information. The method is particularly well suited for the analysis of hierarchically clustered expression profiles. Indeed, all the sets corresponding to all the levels of aggregation are considered. Thus, BlastSets will be able to detect a biological signal wherever it is located in the tree, which confers robustness to the method. This is the case for all information that can be represented through hierarchically aggregated sets such as the physical proximity of genes on the chromosome (Figure 1d).

One of the strengths of this strategy resides in the use of an open-ended unified data representation. As soon as biological information can be attached to sets of sequences, it can be loaded in the database via the standard XML format. Consequently, the BlastSets database can be permanently enriched with new data.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Alain Blanchard, Antoine Danchin, Pascal Durrens, Florian Iragne, Benoit Leblanc and Macha Nikolski for useful discussions and remarks. R.B. has a PhD fellowship provided by the French ministry of Research, N.G. has a PhD fellowship provided by the 'Région Aquitaine'. A.B. was financed by Université Bordeaux 2. The BlastSets project is supported by funds allocated by the ACI IMPBio from the French ministry of Research.

REFERENCES

1. Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.*, **266**, 114–128.
2. Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) Entrez: molecular biology database and retrieval system. *Methods Enzymol.*, **266**, 141–162.
3. Tomita,M., Hashimoto,K., Takahashi,K., Shimizu,T.S., Matsuzaki,Y., Miyoshi,F., Saito,K., Tanida,S., Yugi,K., Venter,J.C. *et al.* (1999) E-CELL: software environment for whole-cell simulation. *Bioinformatics*, **15**, 72–84.
4. de Jong,H., Geiselmann,J., Hernandez,C. and Page,M. (2003) Genetic Network Analyzer: qualitative simulation of genetic regulatory networks. *Bioinformatics*, **19**, 336–344.
5. Danchin,A. (1998) *La barque de Delphes—Ce que révèle le texte des génomes.* Odile Jacob, Paris, France.
6. Nitschke,P., Guerdoux-Jamet,P., Chiapello,H., Faroux,G., Henaut,C., Henaut,A. and Danchin,A. (1998) Indigo: a world-wide-web

review of genomes and gene functions. *FEMS Microbiol. Rev.*, **22**, 207–227.

7. Robinson,M.D., Grigull,J., Mohammad,N. and Hughes,T.R. (2002) FunSpec: a web-based cluster interpreter for yeast. *BMC Bioinformatics*, **3**, 35.
8. Berriz,G.F., King,O.D., Bryant,B., Sander,C. and Roth,F.P. (2003) Characterizing gene sets with FuncAssociate. *Bioinformatics*, **19**, 2502–2504.
9. Gavin,A.C., Bosche,M., Krause,R., Grandi,P., Marzioch,M., Bauer,A., Schultz,J., Rick,J.M., Michon,A.M., Cruciat,C.M. *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.
10. NC-IUBMB. (1992) *Enzyme Nomenclature*. Academic Press, San Diego, CA.
11. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
12. Castillo-Davis,C.I. and Hartl,D.L. (2003) GeneMerge-post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics*, **19**, 891–892.
13. Dufour,J.-M. (1995) *Monte Carlo Tsts with Nuisance Parameters: A General Approach to Finite-Sample Inference and Nonstandard Asymptotics in Econometrics*. Technical Report, C.R.D.E., Université de Montréal, Montreal, Canada.
14. Iragne,F., Barré,A., Goffard,N. and de Daruvar,A. (2004) AliasServer: a web server to handle multiple aliases used to refer to proteins. *Bioinformatics*, in press.
15. Ferea,T.L., Botstein,D., Brown,P.O. and Rosenzweig,R.F. (1999) Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proc. Natl Acad. Sci. USA*, **96**, 9721–9726.
16. Gollub,J., Ball,C.A., Binkley,G., Demeter,J., Finkelstein,D.B., Hebert,J.M., Hernandez-Boussard,T., Jin,H., Kaloper,M., Matese,J.C. *et al*. (2003) The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res*., **31**, 94–96.
17. Sherman,D., Durrens,P., Beyne,E., Nikolski,M. and Souciet,J.L. (2004) Genolevures: comparative genomics and molecular evolution of hemiascomycetous yeasts. *Nucleic Acids Res*., **32**, D315–318.
18. Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*., **28**, 27–30.
19. Jansen,R., Greenbaum,D. and Gerstein,M. (2002) Relating whole-genome expression data with protein-protein interactions. *Genome Res*., **12**, 37–46.
20. Inada,T., Winstall,E., Tarun,S.Z.,Jr, Yates,J.R.,III, Schieltz,D. and Sachs,A.B. (2002) One-step affinity purification of the yeast ribosome and its associated proteins and mRNAs. *RNA*, **8**, 948–958.
21. DeLabre,M.L., Kessl,J., Karamanou,S. and Trumpower,B.L. (2002) RPL29 codes for a non-essential protein of the 60S ribosomal subunit in *Saccharomyces cerevisiae* and exhibits synthetic lethality with mutations in genes for proteins required for subunit coupling. *Biochim. Biophys. Acta*, **1574**, 255–261.
22. Planta,R.J. and Mager,W.H. (1998) The list of cytoplasmic ribosomal proteins of *Saccharomyces cerevisiae*. *Yeast*, **14**, 471–477.
23. Lee,J.C., Henry,B. and Yeh,Y.C. (1983) Binding of proteins from the large ribosomal subunits to 5.8 S rRNA of *Saccharomyces cerevisiae*. *J. Biol. Chem*., **258**, 854–858.